



Fox, Christopher ORCID logoORCID: <https://orcid.org/0000-0003-1931-2058> and Morris, Stephen ORCID logoORCID: <https://orcid.org/0000-0001-6869-8933> (2021) Evaluating outcome-based payment programmes: challenges for evidence-based policy. *Journal of Economic Policy Reform*, 24 (1). pp. 61-77. ISSN 1384-1289

Downloaded from: <https://e-space.mmu.ac.uk/622374/>

Version: Accepted Version

Publisher: Taylor & Francis

DOI: <https://doi.org/10.1080/17487870.2019.1575217>

Please cite the published version

<https://e-space.mmu.ac.uk>

Abstract

We review the state of evaluation within outcome-based commissioning in the United Kingdom. This is the first review to include empirical evaluations of both PbR and SIB programmes. We find a paucity of evaluation and that the quality of evaluations is not high. Moreover, studies tend to conflate the outcomes-based commissioning mechanism with the intervention or services that are funded, and are unable to assess the contribution of these separate elements to impact. Our review also highlights the challenges faced by evaluators in measuring social outcomes. We suggest ways to address these challenges.

Introduction

Outcomes-based commissioning is an important element of the public service reform agenda in the UK and comprises two distinct but related approaches: 'Payment by Results' (Pay for Success) and 'Social Impact Bonds' (Pay for Success financing). A Payment by Results (PbR) contract links payment by the commissioner of a service to outcomes achieved by the contracted provider (Cabinet Office 2011). By making some or all payments contingent on agreed outcomes, PbR supposedly reduces the need for 'micro-management' by the commissioner, encourages innovation and transfers risk from commissioning body to provider (National Audit Office 2015). Typically, where payment by results is used it constitutes only a part of the total contract value (Albertson et al. 2018).

In a PbR contract, provision of services by definition occurs before any results are observed and payment made. Deferred payment may favour some classes of provider (those with capital reserves or access to capital) at the expense of others (those whose constitution places restrictions on the use capital reserves or those that cannot raise capital) (Mulgan et al. 2010). SIBs were developed to address this issue (Social Finance 2009). With a SIB the initial capital investment and upfront service running costs are obtained from an investor, rather than the provider responsible for performance of the contract (Edmiston and Nicholls 2017). To date, investors have typically been social investors who consider both social and financial returns. SIBs can be understood as a class of PbR commissioning associated with the broader 'social investment' movement. The size of the social investment market is

difficult to estimate (OECD 2015) but globally, is likely to run to the tens of millions of dollars (GIIN 2017), of which SIBs are a small component.

Outcomes-based commissioning should be understood as part of broader public service reform. The UK Coalition Government's (2010 – 15) White Paper on *Open Public Services* stated that, "Open commissioning and payment by results are critical to open public services ... Payment by results will build yet more accountability into the system – creating a direct financial incentive to focus on what works, but also encouraging providers to find better ways of delivering services" (Cabinet Office 2011: paras 5.4, 5.16.). Latterly, the focus of this reform agenda has been local services (NAO 2015).

Both PbR and SIBs draw on outcome-based performance management (OBPM) (Lowe and Wilson 2015). They might be seen as New Public Management approaches (Hood 1991), although 'Public Governance' (Halligan et al. 2012) with its emphasis on the role of multiple stakeholders and recognition that social outcomes are emergent properties of complex systems (Lowe and Wilson 2015, Lowe 2017) is argued by some to be more appropriate, particularly for SIBs. Given the need to make precise estimates of impact and be able to attribute impact to an intervention – evaluation designs with high levels of internal validity might be expected to be a feature of PbR programmes (Warner 2013)

Our concern in this paper is the extent and quality of the existing evaluations of programmes or interventions delivered through PbR/SIB-funding mechanisms. We outline some of the challenges that face evaluators in assessing the effectiveness of PbR/SIB-funded interventions. Based on a rapid-evidence-assessment of evaluations of PbR/SIB-funded interventions in the UK, we conclude that there is a paucity of evaluations and those that exist are not of a high standard. Moreover, that evaluation of the intervention on the one hand, and funding mechanism(s) on the other, are often conflated leading to a lack of clarity. We argue that randomized controlled trial designs may have a role to play but need to be adapted to meet the requirements of the sector. Suggestions for how evaluations can be better designed to meet the needs of policymakers and practitioners alike are made.

Evidential challenges

We suggest three inter-linked challenges around impact measurement for SIBs and PbR models:

1. Evaluating the implementation and fidelity of programmes commissioned through SIBs/PbRs.
2. Evaluating the social impact of programmes commissioned through outcomes-based commissioning: these are evaluations restricted to studying outcomes subject to payment by results and used to inform whether payments should be made to investors (in relation to SIBs these are what Social Finance (2016) refer to as evaluation of ‘contractual outcome metrics’) or that might also evaluate wider social outcomes.
3. Evaluating outcomes-based payment systems as policy instruments¹. These evaluations are primarily for the benefit to policy-makers, programme designers and future investors (what Social Finance (2016) term ‘impact evaluation for learning’). To date these have usually been process evaluations, but they could also be impact evaluations that seek to understand whether there is a causal link between the policy instrument and outcomes

These challenges are also pertinent to the wider social investment movement.

Methodological approach

While there have been structured reviews of SIBs (e.g. Fraser et al. 2016), this is the first paper to apply a Rapid Evidence Assessment (Government Social Research Unit 2007) methodology to examine empirical evaluations of PbR/SIB programmes together. The results of the review are reported elsewhere (Albertson et al. 2018). In this paper we review the design and use of PbR and SIB evaluations.

In the next section we describe the review methods. We then set out the findings as they relate to the types of evaluation approach. We conclude with a wider discussion and suggestions for future directions.

¹ Policy instruments are techniques through which the State attempts to achieve its goals and that carry specific concepts of the politics/society relationship (Linder and Peters 1990, Lascoumes and Le Galès 2007)

Review methodology

Search strategy

Our focus was on outcome-based commissioning with a focus on social outcomes. We constructed a list of recent UK PbR contracts. Initially we used a list compiled by the National Audit Office (2015: Figure 4). We added additional programmes sponsored by central government departments. It should be noted, however, that PbR contracts have spread beyond those commissioned by national government - identifying these is beyond the scope of this paper. We also exclude National Health Service PbR programmes. These are a particular arrangement more akin to 'payment by outputs' and fall outside this review. PbR programmes commissioned by the Department for International Development are excluded as these are delivered outside the UK. Through a database maintained by Social Finance² we further identified every live SIB in the UK at April 2017.

Identifying evaluations

A search was undertaken for published evaluations associated with each identified PbR and SIB. This included searching websites associated with programmes, their funders, investors and service providers. A structured search of two databases, chosen for their broad coverage of potentially relevant studies, was undertaken: ASSIA and Web of Science. The search was conducted in June 2017 with no time or language filters base on the search terms:

"social impact bond*" OR "pay for success bond*" OR "pay for success contract*" OR "social outcome invest*" OR "social impact invest*" OR "social invest*" OR "payment by results" OR "pay for success" OR "outcome-based commissioning" OR "outcom-based payment" OR "outcome-based funding" OR "outcome commissioning" OR "commissioning for outcomes" AND evaluat*

Screening

After de-duplication 811 papers were screened on titles and abstracts. Inclusion criteria were:

² <http://www.socialfinance.org.uk/database/>

- Reports on empirical evaluations of PbR and SIB programmes in the UK

Exclusion criteria were as follows, reports/papers:

- Of empirical evaluations of PbR and SIB programmes outside the UK.
- That did not report empirical evaluations
- That reported evaluations of NHS PbR programmes
- That reported ex-ante evaluations.

58 papers were retained. The full text of all retained papers was retrieved and subject to a second sift for relevance using the criteria discussed above. Forty-six papers were retained for full analysis. A summary of the papers and the programmes and more detailed account of the methodology is available on-line³.

Quality

Forty six papers were assessed. The quality of qualitative evaluations were judged using standards drawn-up by the UK government (Spencer et al. 2003). These standards are widely used and many of the evaluations identified were commissioned by the UK government. The design of impact evaluations was assessed using Sherman et al.'s (1998) Scale of Scientific Methods (the Maryland Scale). A number of studies, however, were not primarily evaluations but had evaluative elements. In these cases professional judgement was used to assess methodological rigour.

Description of papers in review

Of 46 papers, 30 relate to PbR programmes, 15 to SIBs and one both a PbR programme and SIB (Ministry of Justice 2014). Four papers were published in peer reviewed journals, the remainder are 'grey literature', often published by UK government departments.

The majority of papers (37) are primarily process evaluations. Eight include a quantitative impact evaluation. Some are interim or supplementary reports and relate to five different programmes, of which one is a SIB⁴.

Findings

A paucity of evaluation

There are only a small number of evaluations of PbR and SIB programmes in the UK. Some quite large PbR programmes were not evaluated. For example, the Transforming Rehabilitation programme, where £3 billion worth of probation services were subject to PbR contracts, has not to our knowledge been evaluated. Most PbR evaluations were process evaluations and there were few impact evaluations of funded interventions or commissioning models. There is also little evidence of fully developed economic evaluations.

This paucity of evaluation is consistent an NAO (2015) report. It concluded that “Most operational PbR schemes have yet to finish so there is not yet enough evidence to evaluate the effectiveness of either individual schemes or the PbR mechanism” (NAO 2015: 6). Evaluations of some PbRs have been cancelled or modified before being completed, or evaluation plans become impractical (Webster 2016).

In cases where SIBs in the UK have paid-out, payment has been based on meeting performance targets, not on results from impact evaluation. For example, in the ‘Its All About Me’ (IAAM) Adoption Bond there was no evaluation of implementation or impact. The Cabinet Office stated:

“This cohort of children is very unlikely to have found a home in the absence of this intervention given the rates of adoption and their characteristics. Therefore we assume that none of the cohort would have been placed without IAAM, and deadweight is therefore nil.”⁵

⁴ Shortly after the review was completed the final impact evaluation for the Peterborough SIB was published (Anders and Dorsett 2017). It is not part of the review.

⁵ https://data.gov.uk/sib_knowledge_box/node/183

The one exception is Peterborough SIB, where the decision not to make payment for outcomes for the first cohort was based on an evaluation comprising a matched comparison group design (Jolliffe and Heddermann 2014, Anders and Dorsett 2017).

The paucity of evaluation is also consistent with other reviews. O’Flynn and Barnett highlight what they see as a paradox within impact investing: that the sector is concerned with “the prioritisation of ‘social impact’ without prioritising ‘impact evidence’” (O’Flynn and Barnett 2017: 3). O’Flynn and Barnett suggest that this is due to cost considerations; the administrative burden placed on the investee; that impact is implicitly assumed and so doesn't need to be measured; and that social outcomes might occur many years after investment. To this we would add: the complexity of designing evaluations that can attribute social outcomes to programmes, often due to the difficulty of finding a comparator; and, debates about methodology within the evaluation sector that can be off-putting to commissioners.

The quality of evaluations is not particularly high

As noted, the majority of papers (37) are process evaluations, using either exclusively qualitative methods or mixed method approaches. The majority of evaluations were reasonably strong on data collection, analysis and reporting. Consistent areas of weakness were:

- Absence of a theory of change, an approach adopted in only a handful of evaluations;
- Sampling, for which there was rarely a clear rationale;
- Discussion of design or methodology and how this relates to evaluation questions; and
- Research ethics, addressed in only a minority of studies.

The quality framework used does not set out weightings for different factors, calling instead for the use of informed judgement, nevertheless, the absence of theory and of clear sampling strategies is of particular concern.

Eight papers included a quantitative impact evaluation. Some are interim or supplementary reports. In one example, a study looked at five different programmes, one of which was a

SIB⁶. Of these, none were randomised experiments considered by many to be the most rigorous of impact evaluation designs (Shadish, et al, 2002). Typically designs involved an intervention and matched comparison groups (e.g. Newton et al. 2014, Nafilyan and Speckesser 2014, Ministry of Justice 2014, Jolliffe and Hedderman 2014). These evaluations relied on administrative data sets that were often of variable quality and difficult to access (see for instance Bewley et al. 2016 and Newton et al. 2014). Some weaker designs with no contemporaneous comparison or control group were found (Ministry of Justice, 2015).

Measuring social outcomes is difficult

Where work was done on measuring social outcomes it was closely tied to the interventions themselves. There was little or no mention of capturing wider social outcomes that might be linked to outcomes-based commissioning. According to Disley et al. (2011), the ability of investors and markets to account for social outcome risk is currently underdeveloped: metrics are unclear; financial markets do not price social value creation; and, consequently, there is a lack of comparable performance information (also ATQ Consultants and Ecorys 2015).

Tools such as rate cards have been created and helped stimulate the development of SIBs (e.g. Griffiths et al. (2016). In some cases, outcomes-based commissioning has stimulated a literature on outcome measurement. In a review of a recent SIB to address loneliness, ATQ Consultants and Ecorys (2016) note that the SIB quantified the costs and benefits of loneliness and put forward a stronger 'case for investment'. Tan et al. (2015) describe the use of cost-benefit analysis to develop outcome metrics and development of bespoke information management systems, but also that performance measurement, outcome payment thresholds and values, are rarely transferable between SIBs.

Evaluations of PbR in the criminal justice sector note challenges in simplistic, single, binary outcome measure for PbR (Foster et al. 2013 and Pearce et al. 2015). In others, metrics proliferate, adding to complexity. According to Wong et al. (2015), the number of metrics in the Local Justice Reinvestment Pilots made it harder for providers to work out what interventions to implement. Gosling (2016: 527) takes a critical view of the challenges

⁶ Shortly after the review was completed the final impact evaluation for the Peterborough SIB was published (Anders and Dorsett 2017). It is referred to later in this paper, but was not part of the review.

inherent in defining and measuring social outcomes and argues, in relation to the use of PbR in a Therapeutic Community: “PbR creates a clear dichotomy between the achievement of a successful outcome and demonstration of a recovery journey.” Several evaluations emphasise the cost implications of complex performance management systems (Lane et al. 2013, DCLG 2014 and DCLG 2015).

Evaluations tend to conflate effects of outcomes-based commissioning with those of interventions commissioned

PbR and SIBs were conceived as more than simply mechanisms to cut costs. Greater efficiency, accountability and innovation were all perceived benefits (Cabinet Office 2011). This reflects in part ideologies around the limits of state activity and role of citizens in service design and delivery. The Coalition Agreement (HM Government 2010) drafted by the Conservative and Liberal Democrat parties in 2010, contained several references to outcome-based commissioning and emphasized the new government’s desire to re-think size and role of the State, and a belief in extending individual and community involvement in tackling social problems:

“This [programme for government] offers the potential to completely recast the relationship between people and the state: citizens empowered; individual opportunity extended; communities coming together to make lives better.” (HM Government 2010: 8)

Thus, interest from the perspective of evaluation extends beyond the impact of services and the interventions funded through them.

However, few of the evaluations in our review consider the effects of PbR or SIBs as a policy instruments. Some partially address this through an analysis of the effect of PbR on the market for a type of service or the effect of monetised incentive structures on organisational and individual performance, or whether outcome-based commissioning fostered innovation⁷. In relation to SIBs, the lack of focus on the policy instrument might in part be because most SIBs do not lead to the innovative, new interventions. More typically

⁷ Detail on these findings is included in Albertson et al. (forthcoming)

they are used to scale up interventions that are already evidence-based (Albertson et al. 2018). In relation to PbR evaluations, another reason is likely to be that most evaluations have been commissioned by a government committed to the use of PbR as a policy instrument.

Discussion

Our review highlights the limited number of impact evaluations of SIB/PbR programmes in the UK and limitations in both scope and quality. Explaining this is not straightforward and requires some conjecture.

Of PbR evaluations identified, all but two were commissioned and published by government departments. Looking back at the 'invitations to tender' for the procurement of these evaluations they are often prescriptive, concentrate on process evaluation and direct evaluators to focus on the programme implemented rather than wider evaluative questions. The National Audit Office (2015) highlighted the lack of evaluation evidence to support PbR and noted that business cases produced by government did not always explain why PbR was chosen. The lack of a clear rationale for PbR makes constructing an evaluation harder (National Audit Office 2015). Many commentators have argued that PbR (and SIBs) are part of an ideological project (e.g. Dowling and Harvie 2014, Dowling 2017), and therefore strong political/ideological prior commitment to PBR explains the paucity of evaluation and impact evaluation in particular.

As noted, most SIB evaluations were funded by government departments. The evaluation of Peterborough SIB was funded by the Ministry of Justice, the evaluation of the Innovation Fund by the Department for Work and Pensions and the London Homelessness SIB the Department for Communities and Local Government. That central government has paid for most evaluations of SIBs is probably an indication of their relatively high cost. Moreover, SIBs pursue a range of social outcomes not all of which are easy to measure. This presents technical challenges for evaluation and political challenges in getting diverse stakeholders to agree evaluation aims. Most SIBs that have paid out to date in the UK have relied on the achievement of performance measures not evaluation results.

While few impact evaluation may have been published to date, there is an extensive debate about *how* to do such evaluations. Looking at the Social Investment field O'Flynn and

Barnett (2017: 12) argue that: “While there is no shortage of methodologies claiming to assess social impact . . . most fall short of really capturing impact in its fullest and significant sense.” They reviewed more than 100 social impact assessment tools, frameworks and methodologies and found that:

“[T]here is a substantial body of literature (peer reviews and grey literature) that describes the steps a social impact evaluation should take (i.e. providing the framework), but with little prescription as to the recommended approach, and even less focus on exact tools or instruments for data collection or analysis – with much left to the discretion of the evaluator or impact investor.” (O’Flynn and Barnett 2017: 12)

Evaluating the implementation and impact of commissioned programmes

Virtually all SIBs in the US have been accompanied by a randomised controlled trial (RCT), whereas in the UK none have (Albertson et al. 2018). Social Finance (2016: 2) characterize the evaluation debate in the sector as:

“increasingly polarized among those that maintain that only randomised control trials (RCTs) will do, and those that advocate less intensive approaches in order to accelerate the market.”

Warner (2013) notes that the evaluation field has moved on from debates over the merits or otherwise of RCTs and many evaluators may wearily roll their eyes at yet another discussion surrounding their legitimacy, yet this issue remains live in the social finance sector. We argue that for an increasing number of researchers the RCT is not considered paradigmatic nor consistent with a particular perspective (for example, Bonell, et al., 2018 argue for RCTs within a 'realist' framework). It is rather a research design which in a number of circumstances offers benefits other approaches do not. RCTs are discussed in a variety of contexts where researchers start from different epistemological positions (Bonell, et. al, 2012; Porter, McConnell, & Reid, 2017; Shadish, et al., 2002) – though the approach is predicated on some understanding of an objective reality and therefore is in essence ‘realist’ in the broad sense, beyond this the choice of an RCT design should be one based on whether it offers advantages in addressing the evaluation question(s) at hand.

RCTs have been used successfully to evaluate a range of social interventions (Greenberg & Shroder, 2004). However, their use is not without controversy (Deaton & Cartwright 2017; Pawson & Tilley 1997). We suggest that RCTs can make a useful contribution to the evaluation of PbR/SIB-funded interventions, with two qualifications: (1) they require mixed-method enhancements to account for the additional complexity of PbR/SIB funding mechanism; and (2) they are no silver bullet and not all PbR/SIB-funded programmes will be amenable to their use. Furthermore, they will rarely enable evaluation of the funding mechanisms themselves, as it is difficult to see how funding arrangements can be practically randomly assigned so that causal effects of the intervention can be separated from that of the funding mechanism. The use of cluster randomized trials may have potential but even here the number of clusters to be assigned and monitored for adherence to the design may be prohibitively large. We discussed below newly emerging approaches to impact evaluation that may prove more fruitful.

RCTs nonetheless address the question of attribution, a question at the heart of outcomes-based commissioning. Results from a well-designed randomized experiment can give confidence to commissioners and investors that an investment has delivered the social outcomes intended. Reliable estimates of impact are also needed to support economic evaluation and to understand whether to invest in particular interventions.

Although we advocate the use of randomized experiments we suggest that their use in relation to outcome-based commissioning can be improved. Developments in other sectors suggest that RCTs are more useful when accompanied with:

- programme theory

Due to the complex nature of many PbR/SIB programmes and the intricate relationships between programme or service components and funding mechanisms, randomized experiments should be implemented in conjunction with the development of programme theory. The most common approach to developing programme theory is the theory of change. It is not uncommon to see programmes with logic models but these rarely extend to address the underlying causal mechanisms or processes that interventions are hypothesized to trigger in service users, practitioners and/or managers. Developing full theories of change can help clarify the expected contribution

of elements of the service or programme and separately the funding mechanisms, as well as how these elements might interact. Furthermore, more rigorous approaches to developing programme theories can help evaluators better understand contextual influences that can be important in determining success or failure, what Cartwright and Hardie (2012) refer to as supporting factors and add to understandings of whether interventions will deliver similar results in different contexts. The proponents of approaches to realist evaluation (Pawson & Tilley, 1997) which utilize similar processes (Blamey and Mackenzie 2007) certainly maintain that such approaches to grounded programme-based theorising are consistent with understanding complexity (Pawson, et al., 2005).

- testing of mediators

Incorporating programme theory into randomized experiments expands the types of questions that can be addressed. As we have noted, rigorous theories of change specify a range of causal processes or mechanisms that programmes might trigger. Clearly then an impact evaluation needs to test for such processes or mechanisms. This is central to understanding not only whether PbR / SIB-funded programmes work but also how they work. Moreover, programme theories bring a wide range of benefits particularly in the face of complex interventions. Fortunately advocates of randomized experiments have for some time advanced methods to test for what are termed mediators, which are understood under certain conditions to shed light on mechanisms (for example Baron & Kenny, 1986 and Imai, et al., 2011). These methods offer some promise but also suffer from limitations and drawbacks (Gerring, 2010; Green, Ha, & Bullock, 2010). Whilst we believe that RCTs explicitly incorporating the measurement of proposed mediating factors are worth pursuing, qualitative approaches are now also advocated as a route to explicating causal mechanisms within the context of randomized experiments (Morris, et al, 2016).

- greater attention to context

Integrating rigorous programme theory into the design of RCTs will naturally lead to a renewed emphasis on the importance of context. Traditionally RCTs have been criticized due to a lack of attention paid to external validity, leading to difficulties for

policy makers in knowing whether results from studies conducted in a particular context hold in their or other circumstances (Cartwright & Hardie, 2012). Moreover, an understanding of context when interpreting results from randomized evaluations is an essential component of 'causal explanation'. This concern is reflected in the development of general reporting standards for randomized trials, such as CONSORT, that require discussion of factors likely to influence external validity (generalizability) of findings (Boutron et al., 2017) and is found in sector specific reporting standards such as EMMIE for crime and justice studies (Sidebottom & Tilley, 2012).

- explicit incorporation of mixed methods

Acknowledging the role of rigorous approaches to programme theorizing as an enhancement to RCTs, and the necessary emphasis on causal processes as well as context, points inevitably to mixed method study designs - RCT that incorporate both quantitative and qualitative elements. The call for the enhancement of experiments with qualitative and indeed other methods has grown apace in recently years, with researchers in health services research field leading the way (Oakley, et al, 2008) but being joined by political scientists (Paluck, 2010), educationalists (Hanley, Chambers, & Haslam, 2016; Maxwell, 2012) and criminologists (Sherman & Strang, 2004). Qualitative methods are often described as being part and parcel of process evaluation, which combines a range of approaches to explore 'how-type-questions' and to add what Collier, Brady, & Seawright, (2010) refer to as 'causal leverage'. Our review of evaluation practice suggests that not only should researchers in the field of PbR and SIB evaluation take RCTs seriously but that they should develop genuinely mixed method studies. We therefore advocate a pragmatic approach to evaluation where selection of methods are made in an explicit manner based on a reasoned judgement as to how effective they might be in addressing evaluation questions (Biesta, 2015; Johnson, Onwuegbuzie, & Turner, 2007).

These enhancements to the basic randomized design can enable experiments to contribute greater levels of insight in cases where interventions are relatively complex, context dependent, and interact in complicated ways with the funding mechanisms. This more 'rounded' approach to RCTs suggests scope for the more effective integration of process and impact evaluation, underpinned programme theory. In the case of SIB/PbR programmes this often requires the evaluation design to be integrated into intervention design. In many

cases it is difficult to introduce randomization *after* an intervention has already been designed. On a practical level this implies that evaluators should be appointed earlier in the process of designing a SIB, but more fundamentally it suggests the need to train policy-makers and commissioners in the tenets of evaluation design.

Of course, randomization is not always possible and where this is the case some of the the quasi-experimental designs that still maintain high levels of internal validity might be appropriate (Cook, Shadish and Wong 2008).

Measuring social outcomes

Notwithstanding the volumous debate around what consistutes social value and how to measure it, we observe that measurement of social outcomes remains a significant challenge and one that was readily apparent from the review we undertook. Debate in the literature and across the sector also focuses on the degree to which general solutions to this problem can be developed, or whether sector-specific approaches are all that can be expected (Rawhouser, Cummings, & Newbert, 2017). The scale of this challenge defies any attempt to discuss the issues at length here. Nonetheless we offer some observations on how the challenges might be addressed from the perspective of practical research design, rather than immerse ourselves in theoretical debates. [here]

First, many of the interventions funded through PbR contracts or SIBs are not particularly novel or innovative. This means that evaluators can learn from the existing evidence and research about how social outcomes of relevance might best be measured. Our review suggests there may be a tendency among evaluators and policymakers in this sector not to engage with the existing research. This may stem in part from the tendency of government to commission evaluations and from studies to be undertaken by consultants rather than academics.

Second, many evaluations for quite obvious reasons, tend to rely on administrative data sets. While this might be a strategy to reduce costs, it may lead to problems in measuring appropriate constructs. Clearly administrative/management data sets are not established with the needs of evaluation in mind, nor do the measures typically derived reflect the concerns of construct validity and measurement theory. Unfashionable though it might be

in an age of big data, we wonder whether the PbR/ social finance sector should revisit the basics of sample and survey design, questionnaire design and testing, and not so readily dismiss primary data collection, noting that online and digital environments open up new possibilities for survey work.

Finally, without wishing to suggest that programme theory represents some form of panacea, our last observation is that approaches such as theories of change offer the potential to better understand what outcomes need to be measured and why. This is particularly relevant to SIBs where it is hard to hold organisations or individuals accountable for delivering social outcomes that are emergent properties of a complex system and where

“Complexity approaches seem to call for a different form of accountability. If we are to hold people accountable for exercising good judgement, we need to be familiar with the context of that judgement, and have a detailed account of the way that judgement was exercised.” (Lowe 2017: 80).

[Evaluating outcomes-based payment systems as policy instruments](#)

Evaluating outcomes-based payment systems as policy instruments (the third methodological challenge we introduced earlier) presents a different set of evaluation challenges.

We agree with Social Finance (2016) that evaluation of outcomes-based payment systems for the benefit of the wider sector may well need to be designed and implemented separately from evaluations of the interventions funded. Social Finance (2016) caution against using outcome metrics designed to inform payment decisions for this form of evaluation. Instead, such broader lessons should come through ‘Impact Evaluations for Learning’, which have no contractual bearing and are designed so as not to compromise effective implementation (ibid.). This raises questions about how best to commission these evaluations. Neither government nor social investors will be well-placed due to potential conflicts of interest. In the UK funding is therefore more likely to come from UK Research and Innovation, which brings together the different UK Research Councils. The model of the ‘what works’ centres in the UK might also be relevant. The What Works Centres collate existing evidence on effective policies and practices, promote evaluation, produce syntheses and systematic reviews of evidence and disseminate this knowledge. The Centres are based,

in part, on the established National Institute for Health and Care Excellence (NICE) and the Educational Endowment Foundation models. Perhaps the time has come to consider a 'social finance' what works centre.

Turning to methodological challenges inherent in evaluating policy instruments, in many cases RCTs or other counterfactual designs will not be appropriate because of the small number of cases available relating to a specific policy instrument. This may seem obvious to evaluators but is often a point lost in policy debates. Also, the questions policymakers or funders have in relation to impact of policy instruments may render experimental designs less informative.

Faced with such challenges, evaluators have developed and adopted alternative approaches to identifying impact, and switched from discussing 'attribution' to what is termed 'contribution', recognising the importance of supporting factors in understanding impact in more complex settings (Mayne, 2012; Stern et al., 2012). Others have rather unhelpfully described these 'alternatives' as small 'n' approaches (White and Phillips 2012). They are, however, not simply 'qualitative' alternatives to 'quantitative' impact evaluation. Their proponents are generally critical of relativist perspectives associated with some researchers working in the qualitative tradition. They propose impact designs that their advocates argue enhance causal leverage in circumstances of complexity and uncertainty by foregrounding participants' perspectives, an understanding of the context, and multiple causes or causal packages that lead to impact.

The starting point for such approaches is often a recognition of the complexity of social programmes that involve partnership approaches and have multiple goals (Pawson and Tilley 1997; Blamey and Mackenzie 2007). Put crudely, these alternative approaches see causation as multiple and 'complex'. Interventions operate in complex systems where:

[T]rajectories and transformations depend on all of the whole, the parts, the interactions among parts and whole, and the interactions of any system with other complex systems among which it is nested and with which it intersects." (Byrne 2009: 2)

Case-based approaches, as one example of these alternatives, are based on generative understandings of causation rather than the statistical counterfactual-based perspective

(Byrne et al. 2009, Stern et al. 2012). Moreover, Advocates of case-based approaches reject the “disembodied variable” of quantitative approaches (Byrne 2009: 4). The case is a complex entity in which multiple causes interact:

It is how these causes interact as a set that allows an understanding of cases

This view does not ignore individual causes of variables but examines them as ‘configurations’ or ‘sets’ in their context. (Stern et al. 2009: 31)

Case-based methods can be broadly typologised as either between case comparisons (such as qualitative comparative analysis) or within case analysis (the obvious example of such an approach being process tracing) (Byrne 2009, Befani and Stedman-Bryce 2016). Generally, a sharp distinction between quantitative and qualitative methods is rejected (Stern et al. 2012). White and Phillips (2012) in their review discuss the merits of a number of small n approaches including General Elimination Methodology; Process Tracing and Contribution Analysis. Such methods might have particular value when exploring the impact of different approaches to outcome-based commissioning and social investment as policy instruments, where cases are few, context is key and interventions are complex.

Conclusion

Returning to the challenges we outlined at the start of this paper our review suggests that, currently, the evidence-base to support PbR/SIB-funded programmes is limited, particularly in relation to the impact of interventions funded through such commissioning models and the impact of outcome-based commissioning as a policy tool. Some of the reasons for this are, for sure, political but the high cost of evaluation plays a part, particularly in relation to SIBs.

However, if we are to make better judgements about whether and when outcome-based commissioning models are appropriate the evidence base must be developed and there is much to learn from innovative evaluations in other fields. There is the potential for more effective impact evaluations, including RCTs, integrated with appropriate qualitative designs that draw on programme theory and foreground context. The overall sequencing of evaluations is important with smaller scale studies preceding larger ones in a ‘test learn adapt’ sequence. As is the development of distinct evaluation strategies for evaluating programmes on the one hand, and evaluation of outcomes-based commissioning as a policy

instrument on the other. This in turn suggests that the future design of PbR programmes should be centred on developing smaller programmes (tens of £millions, rather than £billions) for tightly defined services, accompanied by more detailed, holistic, evaluation and that for SIBs, well-designed impact evaluations should be paramount.

Bibliography

Albertson, K., Bailey, K., Fox, C., LaBarbera, J., O'Leary, C. and Painter, G. (2018) *Payment by Results and Social Impact Bonds: Outcome-based payment systems in the UK and US*, Bristol: Policy Press

Anders, J. and Dorsett, R. (2017) *HMP Peterborough Social Impact Bond - cohort 2 and final cohort impact evaluation*, London: NIESR

ATQ Consultants and Ecorys (2015) *Ways to Wellness Social Impact Bond: The UK's First Health SIB: A Deep Dive Report*. London, Commissioning Better Outcomes Evaluation

ATQ Consultants and Ecorys (2016) *Reconnections Social Impact Bond: reducing loneliness in Worcestershire*, London, Commissioning Better Outcomes Evaluation

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.

Befani, B., & Stedman-Bryce, G. (2016). Process Tracing and Bayesian updating for impact evaluation. *Evaluation*. <http://doi.org/10.1177/1356389016654584>

Bewley, H., George, A., Rienzo, C. & Portes, J. (2016) National Evaluation of the Troubled Families Programme: National Impact Study Report. London: DCLG

Biesta, G. (2010). Pragmatism and the Philosophical Foundations of Mixed Methods Research. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of Mixed Methods in Social & Behavioral Research*. Thousand Oaks, California: SAGE Publications, Inc. <http://doi.org/10.4135/9781506335193>

Blamey, A. and Mackenzie, M. (2007) 'Theories of change and realistic evaluation: peas in a pod or apples and oranges?', *Evaluation* 13: 439–55.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2012). Realist randomised

controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine*, 75, 2299–2306.

Bonell, C., Moore, G., Warren, E., & Moore, L. (2018). Are randomised controlled trials positivist? Reviewing the social science and philosophy literature to assess positivist tendencies of trials of social interventions in public health and health services. *Trials*, 19(1), 238.

Boutron, I., DG, A., Moher, D., KF, S., Ravaud, P., & Group, for the C. N. P. T. (2017). Consort statement for randomized trials of nonpharmacologic treatments: A 2017 update and a consort extension for nonpharmacologic trial abstracts. *Annals of Internal Medicine*, 167(1), 40–47. Retrieved from <http://dx.doi.org/10.7326/M17-0046>

Byrne D (2009) 'Case-based methods: why we need them; what they are; how to do them', in Byrne D and Ragin CC (eds) *The SAGE Handbook of Case-Based Methods*. London: Sage.

Cabinet Office (2011) Open Public Services White Paper, London: Cabinet Office.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Collier, D., Brady, H. E., & Seawright, J. (2010). Sources of leverage in causal inference: Toward an alternative view of methodology. In H. E. Brady & D. Collier (Eds.), *Rethinking social inquiry: Diverse tools, shared standards* (2nd ed., pp. 161–199). Lanham, MD: Rowman and Littlefield.

Cook, T., Shadish, W., & Wong, V. (2008). Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons. *Journal of Policy Analysis and Management*, 27(4), 724-750.

DCLG (2014). *Supporting people payment by results pilots: final evaluation*. London: DCLG.

DCLG (2015). *Qualitative evaluation of the London homelessness social impact bond: Second interim report*. London, DCLG.

Deaton, A., & Cartwright, N. (2017). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*.

<http://doi.org/10.1016/j.socscimed.2017.12.00>

- Disley, E., Rubin, J., Scraggs, E., Burrowes, N., Culley, D. and RAND Europe (2011) *Lessons learned from the planning and early implementation of the Social Impact Bond at HMP Peterborough*, London MoJ.
- Dowling, E. and Harvie, D. (2014) 'Harnessing the Social: State, Crisis and (Big) Society'. *Sociology*, 48(5), 869–886
- Dowling, E. (2017). 'In the wake of austerity: social impact bonds and the financialisation of the welfare state in Britain', *New Political Economy*, 22(3), 294-310
- Dunning, T. (2008). Natural and field experiments: The role of qualitative methods. *Qualitative & Multi-Method Research*, 6(2), 17–23.
- Edmiston, Daniel & Nicholls, Alex. (2017). Social Impact Bonds: The Role of Private Capital in Outcome-Based Commissioning. *Journal of Social Policy*, 1-20
- Foster, R., Small, L., Foster, S., Skrine, O., Hunter, G. & Turnbull, P. (2013) Evaluation of the Employment and Reoffending Pilot: Lessons learnt from the planning and early implementation phase. London: Ministry of Justice.
- Fraser, A., Tan, S., Lagarde, M. and Mays, N. (2016) 'Narratives of Promise, Narratives of Caution: A Review of the Literature on Social Impact Bonds', *Social Policy & Administration* 52(1), 4-28
- Gerring, J. (2010). Causal mechanisms: Yes, but.... *Comparative Political Studies*, 43(11), 1499–1526. <http://doi.org/10.1177/0010414010376911>
- GIIN (2017) Annual Impact Investor Survey, New York, GIIN
- Gosling, H. (2016). 'Payment by results: Challenges and conflicts for the Therapeutic Community', *Criminology & Criminal Justice* 16(5), 519-533.
- Government Social Research Unit (2007) *Background Paper 2 – What do we already know? Harnessing existing research*, London: Cabinet Office
- Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough Already about “Black Box” Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose. *The Annals of the American Academy of Political and Social Science*, 628, 200–208.
- Greenberg, D. H., & Shroder, M. (2004). *The digest of social experiments*. The Urban Institute.

- Griffiths, R., Thomas, A. & Pemberton, A. (2016) Qualitative evaluation of the DWP Innovation Fund: Final Report. London: Department for Work and Pensions.
- Halligan, J., Sarrico, C. and Rhodes, M. L. (2012), On the road to performance governance in the public domain? *International Journal of Productivity and Performance Management*, 61(3), 224–234.
- Hanley, P., Chambers, B., & Haslam, J. (2016). Reassessing RCTs as the 'gold standard': Synergy not separatism in evaluation designs. *International Journal of Research & Method in Education*. <http://doi.org/10.1080/1743727X.2016.1138457>
- HM Government (2010) The Coalition: Our programme for government, London: Cabinet Office
- Hood, C. (1991) A Public Management for all Seasons? *Public Administration*, 69(1), 3-19.
- Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4), 765–789.
- Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, 1(2), 112–133. <http://doi.org/10.1177/1558689806298224>
- Jolliffe, D. and Hedderman, C. (2014) Peterborough Social Impact Bond: Final Report on Cohort 1 Analysis, London: Ministry of Justice
- Lane, P., Foster, R., Gardiner, L., Lanceley, L. & Purvis, A. (2013) Work Programme Evaluation: Procurement, supply chains and implementation of the commissioning model. London: DWP
- Linder, S. and Peters, B. (1990) 'The Design of Instruments for Public Policy.' In Nagel, S. (Ed) *Policy Theory and Policy Evaluation*. Westport, CT: Greenwood Press.
- Lascombes, P. and Le Galès, P. (2007) 'Introduction: Understanding Public Policy through Its Instruments—From the Nature of Instruments to the Sociology of Public Policy Instrumentation', *Governance* 20(1), 1 - 21
- Lowe, T. (2017) Debate: Complexity and the performance of social interventions, *Public Money & Management*, 37:2, 79-80

- Lowe, T and Wilson, R (2015) 'Playing the game of out-comes-based performance management: is gamesmanship inevitable? Evidence from theory and practice', *Social Policy and Administration*, DOI: 10.1111/spol.12205
- Maxwell, J. A. (2012). The Importance of Qualitative Research for Causal Explanation in Education. *Qualitative Inquiry*, 18(8), 655–661.
<http://doi.org/10.1177/1077800412452856>
- Mayne, J. (2012). Contribution analysis: Coming of age? *Evaluation*, 18(3), 270–280.
<http://doi.org/10.1177/1356389012451663>
- Ministry of Justice (2014) *Peterborough Social Impact Bond HMP Doncaster: Payment by Results pilots – Final re-conviction results for cohorts 1*, London: Ministry of Justice
- Ministry of Justice (2015) *HMP Doncaster: Payment by Results pilot – Final re-conviction results for cohort 2*, London: Ministry of Justice
- Morris, S. P., Edovald, T., Lloyd, C., & Kiss, Z. (2016). The importance of specifying and studying causal mechanisms in school-based randomised controlled trials: lessons from two studies of cross-age peer tutoring. *Educational Research and Evaluation*, 22(7–8), 422–439. <http://doi.org/10.1080/13803611.2016.1259113>
- Mulgan, G., Reeder, N., Aylott, M. and Bosher, L. (2010) *Social Impact Investment: The Opportunity and Challenge of Social Impact Bonds*, London: The Young Foundation
- National Audit Office (2015) *Outcome-based payment schemes: government's use of payment by results*, London: NAO.
- Nafilyan, V. and S. Speckesser (2014) The Youth Contract provision for 16- and 17-year-olds not in education, employment or training evaluation: Econometric estimates of programme impacts and net social benefits. London, DfE
- Newton, B., Speckesser, S., Nafilyan, V., Maguire, S., Devins, D. & Bickerstaffe, T. (2014) The Youth Contract for 16-17 year olds not in education, employment or training evaluation: Research Report, London: DfE.
- Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. *British Medical Journal*, 332, 413–415.

- OECD (2015) Social Impact Investment: Building the Evidence Base, Paris: OECD
- O'Flynn, P and Barnett, C. (2017) *Evaluation and Impact Investing: A Review of Methodologies to Assess Social Impact*, Brighton: Institute of Development Studies
- Paluck, E. L. (2010). The promising integration of qualitative methods and field experiments. *The Annals of the American Academy of Political and Social Science*, 628, 59–71.
- Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review-a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy*, 10(1_suppl), 21–34.
- Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage Publications.
- Pearce, S., Murray, D. & Lane, M. (2015) HMP Doncaster Payment by Results pilot: Final process evaluation report. London: MoJ
- Porter, S., McConnell, T., & Reid, J. (2017). The possibility of critical realist randomised controlled trials. *Trials*, 18(1), 133.
- Rawhouser, H., Cummings, M., & Newbert, S. L. (2017). Social impact measurement: Current approaches and future directions for social entrepreneurship research. *Entrepreneurship Theory and Practice*, 1042258717727718.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin and Company.
- Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P. and Bushway, S. (1998) 'Preventing Crime: What Works, What Doesn't, What's Promising', *National Institute of Justice Research in Brief*, Washington D. C.: National Institute of Justice
- Sherman, L. W., & Strang, H. (2004). Experimental Ethnography: The Marriage of Qualitative and Quantitative Research. *The Annals of the American Academy of Political and Social Science*, 595, 204–222. Retrieved from <http://www.jstor.org/stable/4127621>
- Sidebottom, A., & Tilley, N. (2012). Further improving reporting in crime and justice: an addendum to Perry, Weisburd and Hewitt (2010). *Journal of Experimental Criminology*, 8(1), 49–69.

Social Finance (2009) Social Impact Bonds Rethinking finance for social outcomes, London:

Social Finance

Social Finance (2016) Balancing Evidence and Risk, London: Social Finance

Spencer L. Ritchie J, Lewis J and Dillon L (2003) Quality in Qualitative Evaluation: A framework for assessing research evidence, London: Cabinet Office.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations: Report of a study commissioned by the Department for International Development. DFID: Department for International Development.*

Tan, S., Fraser, A., Giacomantonio, C., Kruithof, K., Sim, M., Lagarde, M., Disley, E., Rubin, J. and Mays, N. (2015) An evaluation of Social Impact Bonds in Health and Social Care: Interim Report. London: PIRU

Warner, M. (2013) 'Private finance for public goods: social impact bonds', *Journal of Economic Policy Reform* DOI: 10.1080/17487870.2013.835727

Webster, R. (2016) *Payment by Results: Lessons from the Literature*, www.russellwebster.com [accessed 01-03-16]

White H and Phillips D (2012) *Addressing Attribution of Cause and Effect in Small n Impact Evaluations: Towards an Integrated Framework*, Working Paper 15, International Initiative for Impact Evaluation.

Wong, K., Ellingworth, D. & Meadows, L. (2015) *Local Justice Reinvestment Pilot: Final process evaluation report*. London: MoJ.